

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 09:56:35

PAGE 1

REFERENCE NO: 225

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

Author Names & Affiliations

- James Wilgenbusch - University of Minnesota
- Claudia Neuhauser - University of Minnesota

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Bioinformatics, Genomics, Agriculture, Mathematics, Supercomputing, Advanced Cyberinfrastructure

Title of Submission

Beyond Pillars and Paradigms: Persistent Challenges Related to Computational and Data Intensive Research

Abstract (maximum ~200 words).

Great strides have been made over the past several decades to support demands related to computational and data intensive research. The following text highlights four broad categories of persistent challenges on our campus today. Addressing these challenges, with support of federal funding agencies is imperative to help the Nation more fully realize its investments in campus and national cyberinfrastructures.

- Workflow Bottlenecks: Specialized data acquisition, compute, and storage infrastructures remain isolated. Connecting and sharing resources, even within a single institution, remain extremely challenging for a variety of reasons.
- Data Privacy Obstacles: New public-private partnerships offer exciting research opportunities, but also require new mechanisms to selectively secure data that may be viewed as the intellectual property of a participating partner.
- Data Integration Difficulties: New instruments, computers, and sensors are generating data at an impressive rate. Consolidating, correcting, cataloging, compiling, and storing these data in a way that they can be easily integrated, challenges the limits of our physical systems, software, and budgets.
- Lack of Long Term Data Preservation Facilities: Most universities lack a suitable infrastructure for long-term preservation of valuable research data, together with information on analysis tools needed to redo the analysis.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

There should be no question that the way we do science today has greatly changed over the last couple of decades. Computational approaches are broadly applied in science and engineering, in part because of the widespread availability of effectively managed computational resources. Data intensive analyses are transforming the way we think about research, in part because large scale simulations, social media outlets, Unmanned Aerial Vehicles (UAVs), long-read genome sequencing, high resolution automated imaging instrumentation, and highly sophisticated remote sensors really have delivered incredibly diverse and in some cases massively "Big Data" sets. Therefore, whether one subscribes to the idea that science is buttressed by two or three pillars; or whether one considers data intensive research a fourth paradigm of the scientific method, we should all agree that computational and data intensive research are an increasing part of most scientific endeavors and have changed the size and the shape of the questions being answered by science today.

Against this background, universities have done a laudable job addressing computational and data intensive research challenges. That said, some challenges still persist and new challenges have also emerged. In general, these challenges fall under four broad categories: 1. Workflow Bottlenecks; 2. Data Privacy Obstacles, 3. Data Integration Difficulties, and 4. Lack of Long Term Data Preservation Facilities. These challenges are not restricted to any one particular research domain; rather, they cut across research domains and impact research activities differently depending on a host of factors. In the following section we provide a couple of tangible examples of the research activities at the University of Minnesota that are impacted by the challenges listed above. Addressing these challenges while continuing to provide a high level of support for research domains that have more traditional roots in high performance computing is critical.

Agroinformatics

The Minnesota Supercomputing Institute (MSI) has teamed with researchers in the College of Food, Agricultural and Natural Resource Sciences (CFANS) to harness the power of worldwide agricultural data for the public good from public breeding programs, individual farmers and farmer cooperatives, and multinational companies. The partnership has resulted in the International Agroinformatics Alliance (IAA). The overarching goal of IAA is to accelerate sustainable productivity growth in local and global agriculture through the development, deployment and stewardship of innovations in food and agricultural systems. IAA's primary focus is on the interoperability of data and applications; database building, while important, is secondary, and comes as a consequence of the informatics and analytics toolkit. In particular IAA aims to go beyond conventional G (Genomics) or G x E (Genomics x Environment) interactions to encompass G x E x M x S (Genomics x Environment x Management x Socio-economics) that not only improve crop breeding platforms but also enhance the deployment, stewardship and commercialization of new crop technologies in ways that improve both their economic and environmental outcomes. Key elements of the data platform include: a novel web-accessible database design that facilitates private and public data pooling; authentication protocols that segment and share data in ways that respects and preserves IP ranging from open to proprietary data; control of what is shared with whom is retained by the data originator. There are several primary challenges to combining these disparate data sources in a coherent and meaningful manner. The first is to consolidate sufficient storage and compute cycles to manage, process, and analyze the data. The second, arguably larger challenge is to have the appropriate infrastructure in place to ensure flexible control over data access to partners utilizing the platform. Although many communities have opted to use a completely open data platform, we are cognizant that most commercial entities, and some public entities, are not willing to place all of their hard-earned intellectual property into an open pool, but are prepared to share data with designated research collaborators as enabled by IAA.

Core Genomics and Proteomics Facilities

The University of Minnesota hosts a number of internally and externally funded Institutes, Centers, Programs, and Laboratories that play a critical role in advancing the research mission of the University and in turn help to shape research activities on a regional, national, and global scale. The degree to which the facilities interoperate is critical to research outcomes for numerous research domains. For the purpose of this response, we highlight two core facilities that are emblematic of the consequences and of the challenges associated with integrating these facilities.

The University of Minnesota Genomic Center (UMGC) is the sole, central genomics core facility for the University, including the main Twin Cities campus and system campuses elsewhere in the state. Roughly 70% of projects are provided to over 500 UMN investigators, with the remainder to researchers from outside the University. An example of one of the instruments that the UMGCC operates is the Sequel System, a high-throughput Single Molecule Real-Time (SMRT) sequencer, from the manufacturer Pacific Biosciences (PacBio). SMRT sequencing by PacBio is a third-generation sequencing technology that is quickly providing new insights in life sciences by allowing researchers to sequence large contiguous regions of the genome and full gene transcripts in a single pass. This is important because earlier, short read sequence technologies are unable to correctly assemble parts of the genome that contain numerous repetitive elements. Getting the number and order of nucleotides found in regions with repetitive elements correct is very important for a broad set of biological questions, from phylogenetics to genome engineering. This instrument, like those before it, came with analysis software that is not well suited to run in traditional, bare metal-shared HPC environments. The new software breaks existing workflows, which in some cases requires new

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 09:56:35

PAGE 3

REFERENCE NO: 225

hardware and new systems configurations to match the new instrument's requirements. As with other genome technologies, data sets are large and balancing storage requirements with available resources pushes the limits of what is available locally and nationally.

The Center for Mass Spectrometry and Proteomics (CMSP) provides services spanning the entire spectrum of biological mass spectrometry. Users of CMSP services include UMN researchers from over 30 different departments, as well as numerous external academic institutions and private companies. A major emphasis is proteomic applications, from large-scale discovery-based quantitative proteomic profiling in complex mixtures, to characterization of isolated protein complexes and protein post-translational modifications. Targeted proteomic analysis against known proteins of interest within complex mixtures is also provided as a service. CMSP also offers services for small molecule biological mass spectrometry, including discovery-based and targeted metabolomics. CMSP works closely with the MSI, using MSI infrastructure to meet its extensive needs in the transfer, storage and analysis of generated data.

Partnerships among campus instrumentation and research computing facilities, and private industry are valuable, but they are currently impeded by workflow, data privacy, data integration, and data storage challenges. Data sharing models are often built on an oversimplified, "all open or nothing model", and new instruments tend to break existing workflows. Compounding these difficulties is the fact that data are being accumulating at a rate that makes managing and storing them a major impediment to progress. The next section will highlight some of the ways in which these challenges could be overcome.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

In the same way that we must continue our push for exascale computing, we must also mount an effort to address the need for flexible computing environments at the mesoscale, training and employment opportunities for professionals with expertise in research computing workflows and data management, and data storage facilities capable of accommodating new challenges associated with Big Data research lifecycles. These are, in effect, the gaps in our existing cyberinfrastructure and the areas of deficiency that define what is needed in order to address our current research challenges.

Flexible Computing Infrastructures at the Mesoscale

Workflows associated with campus data acquisition instrumentation often require support from local, regional, and/or national HPC facilities. Some successes in the use of HPC facilities to accelerate workflows associated with data acquisition instrumentation has encouraged the idea that the entire computing and data storage components of research workflows should be subsumed by the HPC facilities. Alternatively, some institutions have drawn a bright line between their HPC facilities and complicated data workflows, suggesting that this gap could be filled by commercial cloud providers. In our experience, neither of these approaches appropriately addresses the problems. HPC facilities are often so tightly geared to serving the needs of high performance computing and batch scheduled job submissions that new computing requirements either don't fit at all into the existing infrastructure, threaten to break the existing infrastructure, or workflow implementations are brittle and threaten to break at the slightest change. While commercial cloud providers can sometimes offer a more flexibility environment, data transfer, data privacy, software licensing, and high hosting costs can make this option unattractive if not impossible.

Somewhere in between these alternatives is a need for campus level computing at the mesoscale. Depending on the organization of the university, such mesoscale computing services could be handled in a number of ways. For example, programs like NSF's Campus Cyberinfrastructure (CC*) required close coordination between research groups and a university's central IT provider. Addressing mesoscale computing needs will likely require similar campus-wide coordination. Solutions will require some degree of virtualization to accommodate diverse software requirements at scale, high speed research networks to facilitate data transfers from remote instrumentation facilities, and highly skilled staff to help define workflow requirements from research applications to appropriate hardware platforms. Clearly, it's not necessary for all of this expertise to reside within a single unit on campus. Institutions should be required to reflect on their strengths and weaknesses and justify how best to advance and sustain workflows given the certainty and pace of changes to come.

Well Trained Professionals

Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 09:56:35

PAGE 4

REFERENCE NO: 225

Effective CyberInfrastructure (CI) requires more than just high-end hardware and software. Leading edge research is increasingly collaborative and also increasingly dependent on highly skilled professionals with a mix of highly technical skills. These highly skilled professionals are an integral part of advanced CIs and are very often a key component to whether a project succeeds or fails. CI professionals often possess a mix of skills somewhere between a systems administrator and a research scientist, which makes it difficult for Human Resources staff to appropriately classify and compensate these individuals.

The Cyberpractitioner Workshop (NSF Award 1546711) held in Washington DC on July 13 and 14, 2016 went a long way to better understanding the roles and technical skills of non-tenured professionals working to support our Nation's research enterprise. Continued efforts to understand the role of these professionals in our research enterprise is needed, as well as ways to promote the training, recruitment, and retention of these professionals.

Facilities for Data Storage

Most modern research is limited by challenges related to collecting, storing, and analyzing Big Data. At the same time, most campus infrastructures for research data storage tend to target the upper end of the performance spectrum with a relatively small amount of the infrastructure designed for backup capacity, primarily for disaster recovery. For the most part, these research storage infrastructures are purposefully designed to maintain a high quality of service for sponsored research projects with a relatively short time horizon (e.g. three to five years). In general, this means that university research storage systems are more costly and less flexible in terms of how they can be accessed from outside. Such a one-size fits all approach to data storage makes it difficult to sustain research data long term and to implement more complicated workflows and data analysis pipelines.

A two-pronged approach is needed in order to address the mounting challenges related to data storage. The first prong involves an intensifying of efforts to provide data management training opportunities to our undergraduate and graduate student populations. Few students are equipped with the skills needed to effectively manage their research data. The consequences of this is that campus, regional, and national data storage resources are used inefficiently and valuable research resources are diverted away from critical research questions. In addition, lack of data management training also compromise the quality of results and in some cases can jeopardize key data privacy requirements.

The second prong of this approach has more to do with the suitability of the underlying hardware platform to address the full spectrum of research activities taking place at a research university. Here again, universities should take a campus-wide view of what is needed and carefully consider which service providers are best suited to deliver these storage services. Storage vendors will be quick to describe solutions to address this problem, but these solutions often fall short of what's required and typically require complete integration with the proposed solution. This can be expensive and can limit an institutions ability to effectively collaborate with other institutions who may not have adopted a vendor specific solution. Much more work is required to promote a healthy set of technologies to support a range of storage needs from high performance to highly sharable.

Question 3 Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

NA

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."